

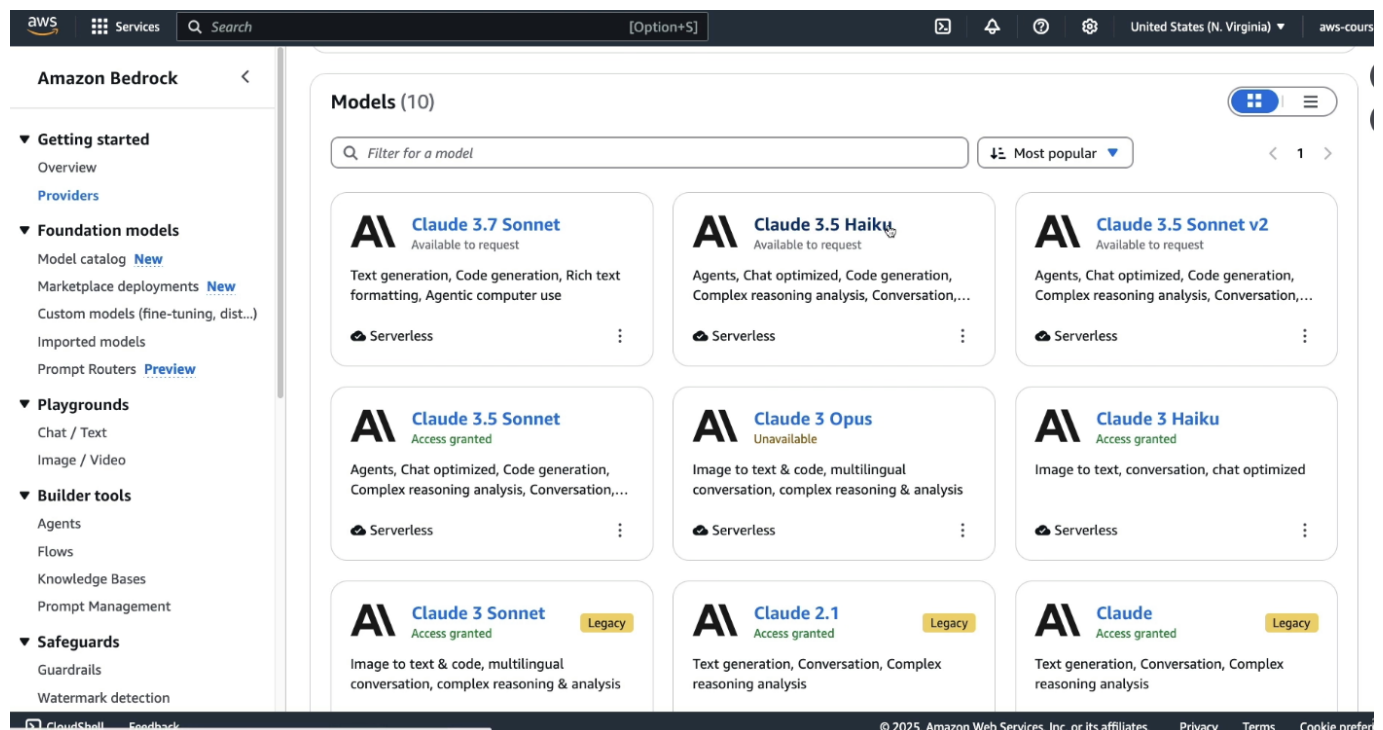
Now that we've looked at some providers, let's go a little deeper.

Different companies and providers offer different models, and each of these models comes with different capabilities. For example, if you choose **Anthropic**, **Amazon**, **DeepSeek**, or **Stability AI**, you'll find that the capabilities vary.

You're **not** expected to memorize which model does what, but the exam will ask you to understand **what a model can and cannot perform**.

For Example:

- If you're looking at a Claude 3.5 Haiku:



- This model is good model for text related tasks:

Amazon Bedrock

Overview
Claude 3 Haiku is Anthropic's fastest, most compact model for near-instant responsiveness. It answers simple queries and requests with speed. Customers will be able to build seamless AI experiences that mimic human interactions. Claude 3 Haiku can process images and return text outputs, and features a 200K context window.

Details

Sold by	Anthropic
Categories	<ul style="list-style-type: none"> Agents Chat optimized Code generation Complex reasoning analysis Conversation Math Multilingual support Question answering RAG Text generation Text summarization Text-to-text Translation
Last version	v1
Release date	Tue, 22 Oct 2024 08:00:00 GMT
Model ID	anthropic.claude-3-5-haiku-20241022-v1:0
Modality	TEXT
Max tokens	200K

On this page
Overview
Usage

© 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

But if you look at a model from **Amazon** like **Nova Reel**:

Amazon Bedrock > **Model catalog** > **Nova Reel**

Nova Reel
By: Amazon | Access granted

Nova Reel is a video generation model. It generates short high-definition videos, up to 9 seconds long from input images or a natural language prompt.

Overview
Nova Reel is a video generation model. It generates short high-definition videos, up to 9 seconds long from input images or a natural language prompt.

Details

Sold by	Amazon
Categories	<ul style="list-style-type: none"> Text-to-video Image-to-video
Last version	v1
Release date	Tue, 03 Dec 2024 08:00:00 GMT
Model ID	amazon.nova-reel-v1:0
Modality	VIDEO
Max tokens	512
Language	English

On this page
Overview
Usage

© 2025, Amazon Web Services, Inc. or its affiliates. Privacy Terms Cookie preferences

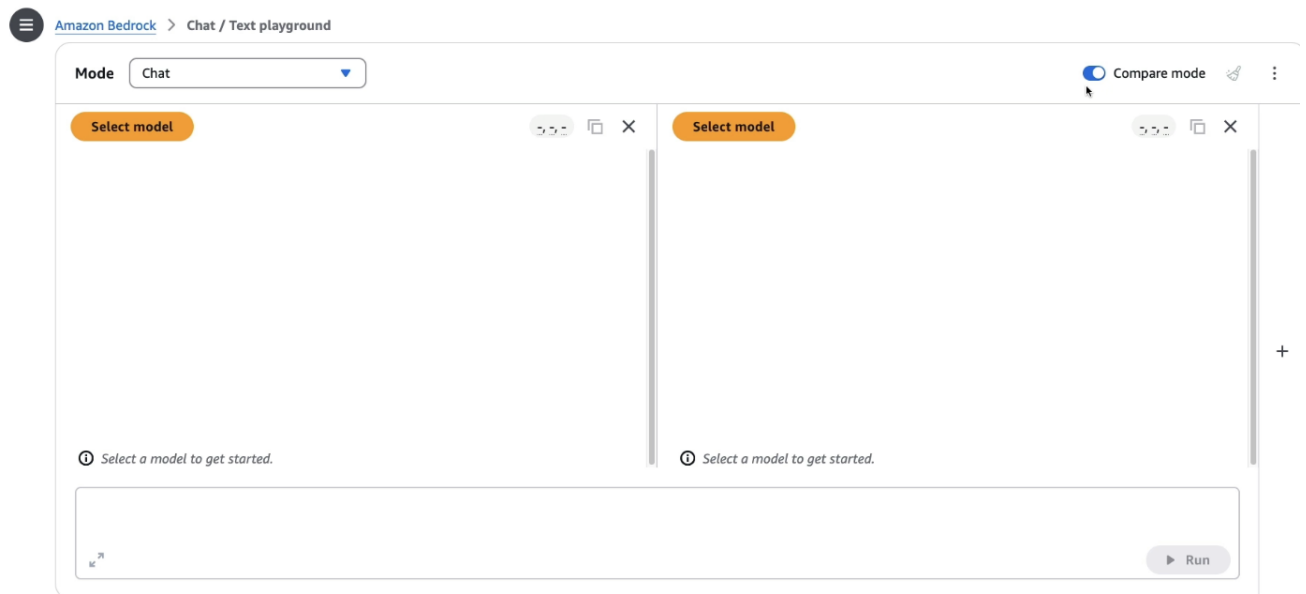
- this model is going to be used for **text to video** or **image to video**.

Each model serves a different purpose. You won't be asked which is "better," but you should be aware of their **capabilities**.

Comparing Models Side by Side

Let's look at how to compare two models:

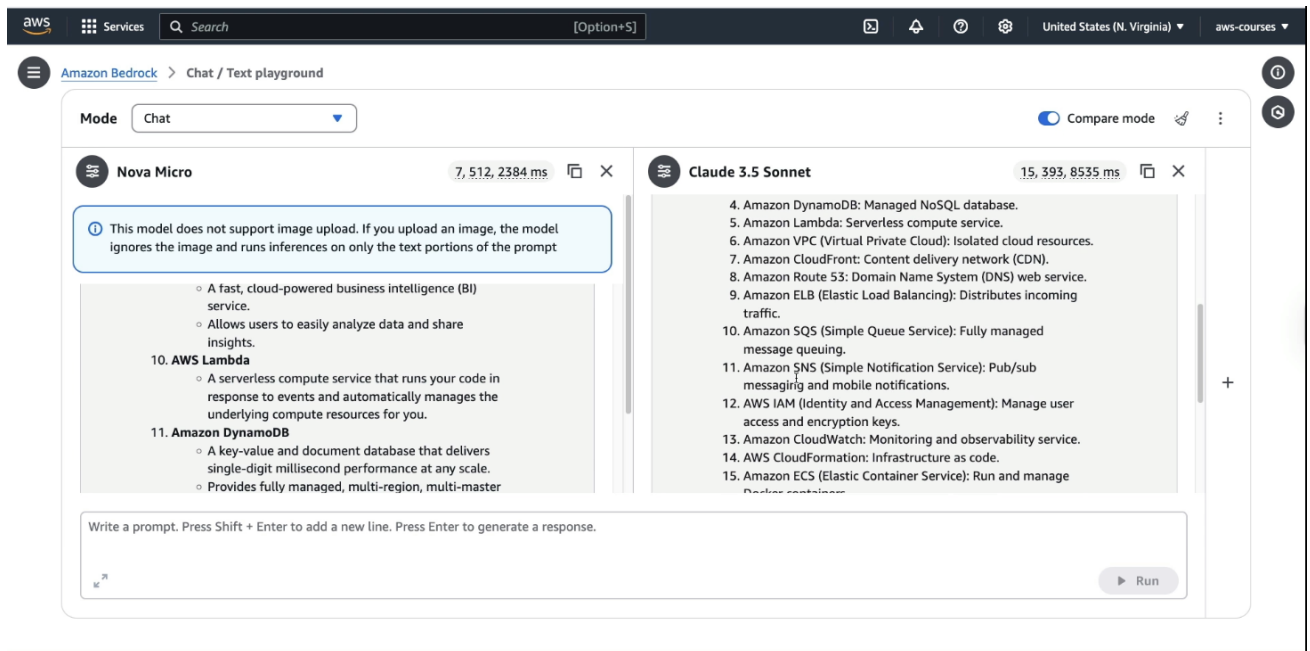
1. **Open the Chats or Text Playground** (I have mentioned it in previous Hands-on how to go about this)
2. **Enable Compare Mode** (top right corner)



3. **Choose two models**, for example:
 - Amazon's **Nova Micro**
 - Anthropic's **Claude 3.5 Sonnet**

Model Differences Observed:

- **Nova Micro** does not support **image upload** – uploaded images will be ignored.
- **Claude 3.5 Sonnet** is **more expensive** than **Nova Micro** but may offer specific benefits for certain use cases.
- Now ask this as a prompt "What are the top AWS services?", the responses vary: (see the image below)



- **Response formatting** and **answer length** differ across models.
- **Performance metrics** (input/output tokens and response time) are visible for comparison.

Example:

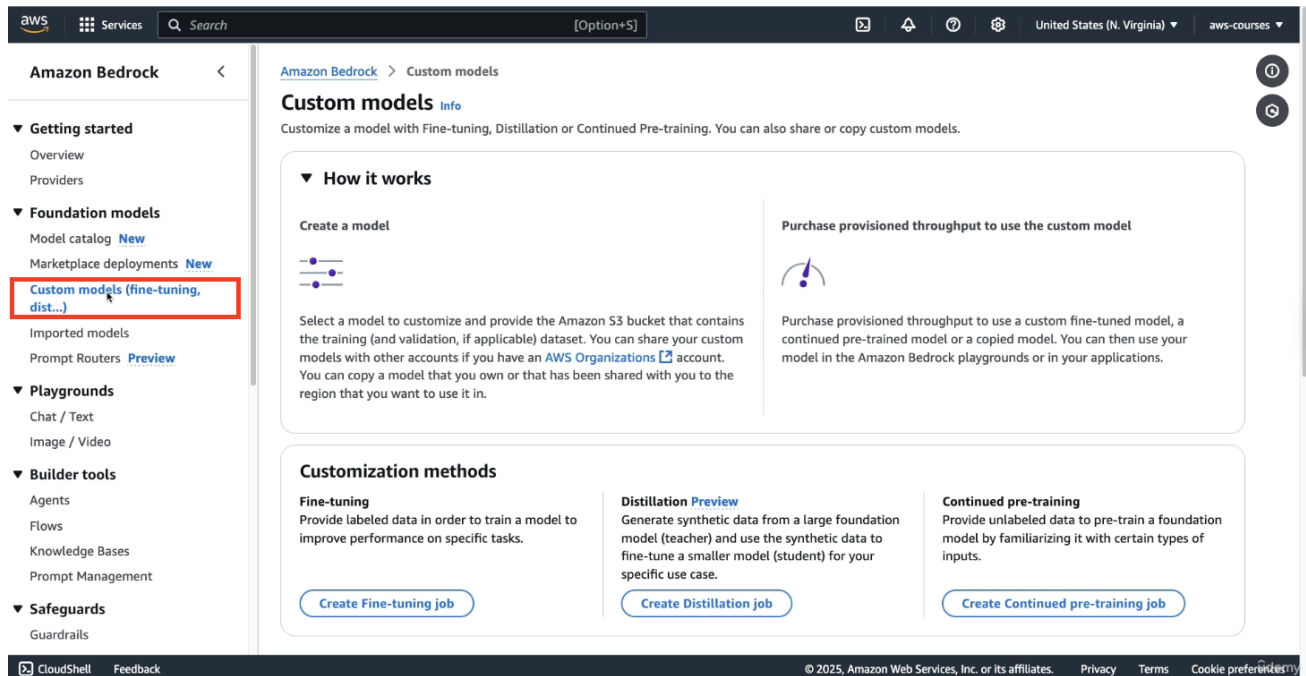
- Nova Micro: 7 input tokens, 512 output tokens, faster.
- Claude 3.5 Sonnet: 15 input tokens, 393 output tokens, slower.

The **takeaway**: consider **output quality**, **response time**, and **cost** when selecting a model.

Exploring Custom Models in Bedrock

To customize a model for your use case:

1. Click on **Custom Model** from the left-hand panel.



here we can create a model that's going to be better fitted for our use case.

2. You'll see **three customization methods**:

- **Fine-tuning** – here we provide labeled data in order to train a model to improve the performance on a specific task. For example: you train it and you fine-tune it to say, "Hey, for this kind of question, I want this kind of answer."
- **Distillation** – (Not covered in this session)
- **Continued Pre-training** – Feed new data to the model to expand its knowledge.

Creating a Fine-Tuning Job (Walkthrough)

Let's simulate setting up a fine-tuning job:

1. Click on **Create a Fine-tuning job**

Select a model to customize and provide the Amazon S3 bucket that contains the training (and validation, if applicable) dataset. You can share your custom models with other accounts if you have an [AWS Organizations](#) account. You can copy a model that you own or that has been shared with you to the region that you want to use it in.

Purchase provisioned throughput to use a custom fine-tuned model, a continued pre-trained model or a copied model. You can then use your model in the Amazon Bedrock playgrounds or in your applications.

Customization methods

Fine-tuning
Provide labeled data in order to train a model to improve performance on specific tasks.

Create Fine-tuning job

Distillation Preview
Generate synthetic data from a large foundation model (teacher) and use the synthetic data to fine-tune a smaller model (student) for your specific use case.

Create Distillation job

Continued pre-training
Provide unlabeled data to pre-train a foundation model by familiarizing it with certain types of inputs.

Create Continued pre-training job

ModelsJobs

Models (0)

A list of models that you've customized or that you've copied from other regions or accounts.

< 1 > ⚙

Custom model name	Source	Type	Share status	Creation time
No custom models				

There are currently no resources.

© 2025, Amazon Web Services, Inc. or its affiliates. [Privacy](#) [Terms](#) [Cookie preferences](#)

2. Select a base model, e.g., Nova Micro

3. Enter a name for the model like `DemoCustomModel1` and enter the **job name** like `Jobname`

4. Configure VPC settings (we skip this here)

5. Set **Input Data** location in Amazon S3 (**V Important**)

Input data [Info](#)

Choose a file in the S3 location. The files you choose must be in the [dataset format](#) that the model needs for training. You can check the data format for your specified model for any potential errors using simple python [script](#). You can also use SageMaker Ground Truth to create and label training datasets. [Learn more](#)

S3 location

View [↗](#)

Browse S3

Validation dataset S3 location (optional)

View [↗](#)

Browse S3

Hyperparameters [Info](#)

Epochs
The total number of iterations of all the training data in one cycle for training the model.

⌵ ⌶

Enter an integer between 1 and 5.

Batch size

6. Set **Validation Dataset** (optional, but useful for checking model accuracy)
7. Configure **Hyperparameters** (advanced – skipped for now)

Hyperparameters [Info](#)

Epochs
The total number of iterations of all the training data in one cycle for training the model.

2

Enter an integer between 1 and 5.

Batch size
The number of samples processed before model parameters are updated.

1

Learning rate
The rate at which model parameters are updated after each batch of training data.

0.00001

Enter a float value between 0.000001 and 0.0001

Learning rate warmup steps
Number of iterations over which learning rate is gradually increased to the initial rate specified.

10

Enter an integer between 0 and 20.

the idea of hyperparameters is that you can tweak these parameters to ensure that the custom model is going to be trained the way you want with the performance you want.

8. Set **Output Data** location

Output data [Info](#)

Choose S3 location to store the model validation outputs.

S3 location

[View](#)

[Browse S3](#)

Service access [Info](#)



Bedrock model customization job requires permissions to write to S3 on your behalf.

Choose a method to authorize Bedrock

- ☒ Use an existing service role
- ☐ Create and use a new service role

Service role

[i](#) Purchase provisioned throughput to use fine-tuned model

After this custom model is created, you need to purchase provisioned throughput to be able to use this model.

[Learn more](#)

[Cancel](#)

[Create Fine-tuning job](#)

Output data is where to store the model validation outputs, if you provided a validation input dataset.

- Assign a **Service Role** to allow Bedrock to access S3 and to get this data for the training as well as for the validation

After setup:

- Click **Create Fine-Tuning Job**

Important Notes About Custom Models

- Custom models require **provisioned throughput** for training and usage.
- This is not an **on-demand** process—resources are allocated specifically for you.
- Once created, you'll use **provisioned throughput** again for running inference with the custom model.

(Note this section is just for understanding purpose only)

What is Provisioned Throughput (the concept):

Provisioned throughput means **you are reserving dedicated compute resources** in advance —**not shared or on-demand**. This is like booking a private room instead of using a shared coworking space.

Here’s how it applies:

- To **train or fine-tune** a foundation model (like Amazon’s Nova or Titan), you can’t just run it like other models.
- You must **pre-purchase provisioned throughput**—this means Amazon sets aside specific compute resources for **your** training job.
- This is required because training is a heavy task and can’t be done on shared on-demand servers.

After the Model is Trained (Inference Phase)

- Once your **custom model is ready**, you might think you can now use it like any other model (like Claude, Mistral, etc.).
- But **you still need provisioned throughput again to run predictions/inference** using that model.
- This is because it's **your private model**, and AWS keeps it on dedicated servers just for you.

Phase	What You Need
Training (fine-tuning)	Buy provisioned throughput
Inference (using the model)	Buy provisioned throughput again